Automated Generation of Latent Topics on Emerging Technologies from YouTube Video Content

Clinton Daniel
University of South Florida
cedanie2@usf.edu

Kaushik Dutta
University of South Florida
duttak@usf.edu

Abstract

Topic modeling has been widely adopted by researchers for a variety of different research problems that involve the mining of text corpora to generate a latent set of topics. Specifically, the Latent Dirichlet Allocation (LDA) algorithm is well documented within academic literature in terms of its application and automated topic generation from data sources such as blogs, social media, and other text collections. YouTube now offers access to over a billion auto-generated video transcript documents that have been recorded and posted to its social platform. The availability of this data offers an opportunity for researchers to investigate a variety of topics that are being discussed and posted to the platform. Specifically, we will study, using the LDA algorithm, discussions related to emerging technologies that have been posted on YouTube to better understand what latent topics can be auto-generated and what kind of methodology can be used to analyze this data.

1. Introduction

Emerging technologies often saturate industry with new terms that generate hype about innovative ways of conducting business [15, 16]. These terms have a way of emerging within various industries through a variety of communication means. Perhaps a term was first exposed to industry through a ground breaking academic research article? Sometimes a term becomes a new buzzword, such as "cloud", "internet of things", or "big data", through the introduction by a key note speaker at a large conference. Eventually a collection of terms become a bag of words that can be used to tell a story about an emerging technology. For instance, a bag of words such as ["business", "technology", "computing", "systems", "organization", "people", "time", "need", "information"] might tell the story of an emerging technology that plays an important role in binding the computing or systems needs of people within organizations, the timeliness of technology, and the information driving the business. Due to the continuous and rapid changes in technology, managers, such as CIOs and CTOs, in business are challenged with identifying the story that's being told [14, 15, 16].

Researchers have attempted to address the problem of identifying emerging technologies through the analysis of text by data mining research proposals, publications [1], and patent systems [2]. These research studies demonstrate that emergence of technology can be detected by analyzing the links between clustered structures of words or terms over slices of time. As the clusters observed across time slices begin to demonstrate an increase in quantitative measures, such as the number of associated papers or patents, the technology is then identified as emerging. Although these research studies demonstrate impressive findings for discovering the emergence of technology through innovative methods, they also cite the need to improve the results of their analysis due to the limited availability of data.

Another means of discovery that can be used to acquire the knowledge of what terms are surfacing in emerging technologies include hype cycles [17, 18]. Technological hype cycle reports are generally published by consulting firms with the goal of informing the industry the current trend of specific technologies. Additionally, these reports often define the term used to describe the emerging technology. For instance, the 2016 Gartner "Hype Cycle for Emerging Technologies" report defines the term "IoT Platform" as *software that facilitates operations involving IoT endpoints and enterprise resources* [3]. Furthermore, in this report, the analysts shape the trend of an emerging technology by plotting its term in a curve that illustrates its expectations for mainstream adoption over time. A study written by Lente, Spitters, and Peine looked at evaluating three different technological hype cycles to develop a theory which may explain the differences between the shapes found in each of their visualizations [4]. The analysis in this study concluded that the three hype patterns differed when comparing three different cases of emerging technologies across all three hype cycles. Additionally, the researchers recognized that more knowledge is required to better understand how to

URI: http://hdl.handle.net/10125/50109
ISBN: 978-0-9981331-1-9

effectively use a hype cycle as a resource for actors who are currently involved in an innovative process.

The limited availability of data reported by researchers attempting to identify emerging technologies using data mining methods, as cited in the Cozzens, et al. [1] and Breitzman, et al. [2] studies, and the differences in hype patterns seen between industry reports, such as hype cycles cited in the Walker, et al. study [3], has identified the need to explore new ways of discovering the pattern of terms associated with emerging technologies. On February 16, 2017 YouTube announced in its official blog that it has automatically captioned over 1 billion videos [5]. These captions are made possible with a combination of Google's automatic speech recognition (ASR) technology and YouTube's caption system. The original intent for captioning YouTube videos was to provide more accessible content for the hearing impaired. In addition to videos displaying closed captioning, YouTube offers an exported transcript of the closed captioning text. This data is potentially valuable to understand or discover terms when trying to explore an emerging technology that may have been discussed in YouTube video content. A bag of words can be assembled from this data which could perhaps establish a structure to be analyzed and tell the story of an emerging technology. This YouTube transcript data offers an innovative opportunity to discovering new knowledge and perhaps a better approach to understanding emerging technology terms buzzing within industry.

In this paper, we will demonstrate the use of an automated method that can be used to generate terms that are associated with emerging technology discussions, such as key note speeches at conferences or interviews of technology leaders, transcribed in YouTube videos. These collections of terms will be generated by the Latent Dirichlet Allocation (LDA) algorithm to form structures of topics. These topics will be analyzed through interpretation and visualization with the intent of telling a story about an emerging technology. The results of this analysis will demonstrate how this research method can be used to verify a valid emerging technology topic generated by the LDA algorithm using YouTube video content.

2. Related Work

In a study conducted by AlSumait, Barbara, Gentle, and Domeniconi, researchers analyzed the significance of ranking topics generated by the LDA algorithm [12]. In this study, the researchers confronted the problem of LDA generating insignificant or meaningless topics by defining a set of decision criteria that measure the distance of a topic from a common insignificant description. This was accomplished by developing an automated unsupervised method of analysis for LDA models. This research serves as an excellent example of the technical and potential difficulties involved in automating the analysis of topic created by the LDA algorithm.

A more recent article published in 2017 recognized the need for researchers to compare different topic models and automate the criteria for choosing the best model that provides the best set of topics [9]. In this study, the researchers develop a methodology that involves searching for specific topics as defined by key search words followed by evaluating the quality of the topics generated by the model. Additionally, the method demonstrates a metric that can be used to potentially reflect the human judgement on the generated topic. The researchers concluded their study by recognizing that their metric needed further work to measure the overall quality of a topic model to identify optimal parameters for the LDA algorithm to include number of topics, distribution of topic vectors, and distribution of words in topics.

In a research study titled, "Videopedia: Lecture Video Recommendation for Educational Blogs Using Topic Modeling" [10], researchers designed a system, called Videopedia, that integrates both text-based blogs and online videos. The system then automatically recommends relevant videos for explaining the concepts given in a specific blog. Additionally, the researchers used topic modeling to map the text data found in YouTube video transcripts combined with the text extracted from blogs to create a common semantic space of topics. The LDA algorithm was used to auto generate the topics to be used as input by their recommendation system. During the process of their research, the researchers recognized that the task of finding videos with a suitable correlation with webpage blogs was a difficult task. However, the researchers were able to demonstrate the proper matching between the text found within the blogs and the content within the video transcripts. They concluded their research by recognizing that a topic modeling algorithm, such as LDA, can be effectively used to support a recommendation of video content to the users of their Videopedia system.

3. Research Method Approach

Discovering topics within a collection of documents typically involves the estimation of latent topics for a given corpus using a topic modeling algorithm such as Latent Dirichlet Allocation (LDA). Blei describes the LDA algorithm as, "*a generative probabilistic model for collections of discrete data such as text corpora*" [6].

The LDA algorithm will be applied to a collection of YouTube video transcripts that were gathered using an innovative data collection and analysis research method. This method involves the collection of YouTube video transcripts and their metadata followed by loading each individual document text into a relational database. Once the text data has been successfully loaded into a relational database, the video transcript data can be analyzed and joined to corresponding metadata with the goal of gathering a targeted data set for export. The targeted dataset is identified by a preliminary analysis based on a specific research question. Then, the video transcript's text will be exported as a collection of separate text files from the relational database to a target directory to form the corpus. The corpus can then be evaluated with the LDA algorithm using an R topic modeling package. R will then be used to visualize the results of the LDA algorithm for further analysis.

Table 1: Research method step summary

Step	Description
(1)	Screen YouTube video content for collection
(2.1)	Extract video metadata
(2.2)	Extract captioning transcript data
(3)	Load data into relational database for analysis
(4)	Process all data and export to text corpus
(5)	Topic modeling and analysis from LDA output in R

Table 1 summarizes the steps involved in the research method while Figure 1 illustrates the technology involved throughout the process. Section 3.1 will explain in detail each step involved in the proposed research method.

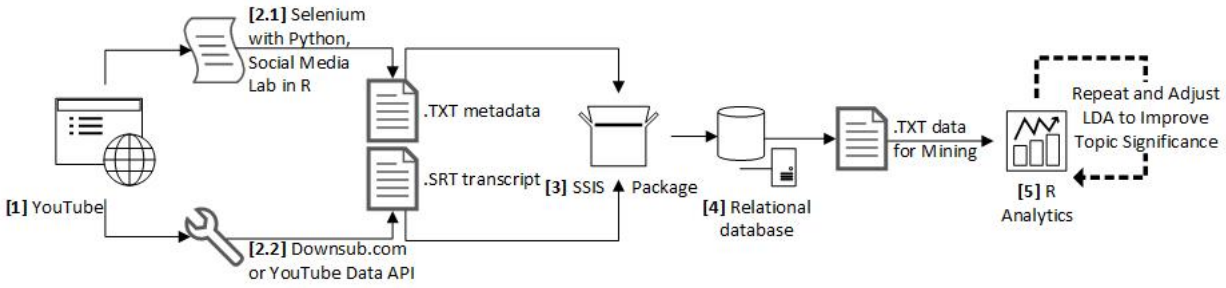

Figure 1: Research method technology summary

3.1. Detailed Steps for Research Method

Step 1 in this research method starts with a search term being used in YouTube's search engine to first find a list of YouTube videos that meet a subject of interest based on a research question. For instance, "emerging technologies" could be used as a search term. Next, the researcher would narrow down the results of the search by reviewing the metadata available from a specific YouTube video of interest. Once a video of interest has been identified by the researcher, the URL will need to be copied to a temporary location for a retrieval at a later time. The following is a sample of the required format to extract the YouTube videos transcript data: https://www.youtube.com/watch?v=msPKD999l7Q.
The most important part of the URL is the video ID. The video ID is identified as the string after the "v=" portion of the URL. For instance, the video ID of this URL is "msPKD999l7Q".

Step 2.1 involves Selenium with Python [7] programming that will be used to extract the associated YouTube video metadata. A custom Selenium with Python program will be used to extract the metadata and save it in a local text (.TXT) file. Some data will be generated and appended to the metadata by the Selenium with Python script to provide further information that is not otherwise provided by YouTube. For instance, the video ID, video URL, date and time of extraction, and YouTube search terms will be added to the generated text file. Additional metadata can be collected, such as comments, using the Social Media Lab package in R. All of this additional information will be used for analysis at a later time. The extracted metadata text file is in a semi-structure format and will be processed in a structured format by another process at a later time in this research method.

A tool is required in step 2.2 to extract the YouTube video captioning transcript. The tool can be custom programmed with a programming language such as Python using Google's YouTube Data API or there are free web-based tools available for data extraction. A web-based tool available at http://downsub.com/ will be used to study this research method and extract the transcript data. After navigating to http://downsub.com/ the researcher will need to enter the YouTube URL of interest in the "Download" tool. This process can be automated using Selenium with Python to re-create the steps needed to acquire the YouTube transcript.

However, the manual steps are documented here for the purposes of understanding the manual process. After clicking the "Download" button, DownSub will extract the transcript data from the YouTube video and make it available for download to your local computer as a .SRT file. .SRT files, or SubRip files, contain the recorded subtitles and timings associated with the specific YouTube video. SubRip files can be opened with text editors to view the recorded subtitles and timings. The SubRip file extracted from DownSub is not in an optimal format for analysis because it contains HTML tags and other information that we are not interested in for analyzing the text at a later time.

In step 3, both the captioned transcript (.SRT) and metadata (.TXT) files will need to be imported into a structured SQL Server relational database. A Microsoft SQL Server Integration Services (SSIS) package will be developed to extract and load the data from the two files into the Microsoft SQL Server database "staging" tables. Data is "staged" in tables within the database so that it can be transformed at a later step in this research method.

Step 4 includes the additional steps added to the Microsoft SSIS package that will be used to transform and process the data from staging tables to the two final structured tables. The structured tables are designed for ease of research analysis. These two tables, named "final_metadata" and "final_transcripts", are described as follows:

Table 2: final metadata relational database table

Column Name	Column Description
Id	Primary key
video_id	YouTube video id
search_terms	YouTube search term to find video
video_time_transcribed	Length of video in seconds
video_title	Title of YouTube video
video_category	YouTube category
subscribe	Number of subscribers for the YouTube channel
views	Number of views of video
published	Date video was published to YouTube
description	YouTube video description
youtube_channel	YouTube channel

Table 3: final transcript relational database table

Column Name	Column Description
id	Primary key
video_id	YouTube video id
transcript	Complete text from transcript

A preliminary analysis can be performed on the final transcript text and metadata tables to determine the value of the data. Once the researcher has evaluated the data, the data can be processed and exported as .TXT files into a targeted local directory for further analysis by R.

Finally, in step 5, topic modeling is performed using the LDA algorithm in R. The LDA algorithm is a generative probabilistic model of a corpus where documents are represented as random mixtures over latent topics. Additionally, each topic is characterized by a distribution over words [6]. The LDA algorithm assumes the following generative process for each document in a corpus [8]:

Algorithm 1: Latent Dirichlet Allocation (LDA)

Step 1: For K topics, choose each topic distribution β_k (Each β_k is a distribution over the vocabulary)

Step 2: For each document in the collection:

a. Choose a distribution over topics θ_d (The variable θ_d is a distribution over K elements)

b. For each word in the document

i. Choose a topic assignment z_n from θ_d (Each z_n is a number from 1 to K)

ii. Choose a word w_n from the topic distribution β_{z_n} (Notation β_{z_n} selects the z_nth topic from Step 1)

The R package 'tm', a framework for text mining applications within R, will be used to generate the objects required to mine the text included in the documents. The R package 'topicmodels' will be used to access and execute the LDA algorithm using the YouTube text data as input. The R package 'reshape2' will be used to restructure and aggregate the data after the LDA algorithm has been executed and prior to visualization of the results. Several R packages, such as 'ggplot2', will be used to visualize the data for analysis. Finally, parameters for the LDA algorithm (such as the number of topics or words generated) will be adjusted as needed to improve the significance of the generated topics.

4. Results and Discussion: Topics on Emerging Technologies in YouTube

To demonstrate how this research method works to discover latent topics within YouTube, we have collected a corpus of 30 transcript documents from YouTube with a search criterion for videos published between 2015 and 2016 that have the terms "emerging technologies" found within their meta data. We will summarize the results of several significant findings in this study to include how this research method can be used to identify valid topics generated by the LDA

algorithm that significantly represent the distribution of documents contained in the corpus. Finally, we will discuss our findings of how this research method can be used to compare the LDA output of a topic generated by YouTube content with an industry hype cycle report that defines the current trends in emerging technologies.

Using the LDA algorithm to discover topics within a document corpus not only creates a distribution of topics populated with words, but it also includes the distribution of topics over documents. Therefore, it was important that we analyze and observe the probability of documents associated with each topic in our corpus. This helped us target a specific set of documents which reside in the corpus that required closer review. To accomplish this, we first studied the collection of 30 documents by breaking them in to two samples. One sample of 15 included only those documents published in 2015 while the other sample included those documents published in 2016. Additionally, the parameter of *K* (number of topics) for the LDA algorithm was reduced to 2 to better understand how an adjustment of *K* affects the potential for an output of valid topics. Overall, this reduced sample size of 30 documents and value for *K* helped us understand how this research method could be used to identify the validity of topics generated by the LDA algorithm. Topic validity was identified by visualizing how strong the topic represented the distribution of documents included in each corpus (the 2015 and 2016 collections). If the topic represented a significant positive or negative linear relationship with the population of documents included the corpus, then the topic was defined as valid. The easiest way to visualize the relationship was to build a scatter plot diagram for all document to topic probabilities included in each corpus.

The following Figures 2 and 3 illustrate the estimated proportions of words from the document population (2015 and 2016) that are generated by a specific topic. The results are illustrated in the form of a scatter plot to visualize the correlation between the two variables (Documents and Topics). For instance, the LDA algorithm had estimated that 76% of the words in YouTube Document 1 were generated from Topic 1 (See Table 4).

Table 4: Document to Topic Probabilities based on 2015 YouTube data sample

Document	Topic 1	Topic2
1	0.76	0.24
2	0.72	0.28
3	0.76	0.24
4	0.26	0.74
5	0.35	0.65
6	0.72	0.28
7	0.79	0.21
8	0.08	0.92
9	0.76	0.24
10	0.67	0.33
11	0.80	0.20
12	0.48	0.52
13	0.75	0.25
14	0.66	0.34
15	0.60	0.40

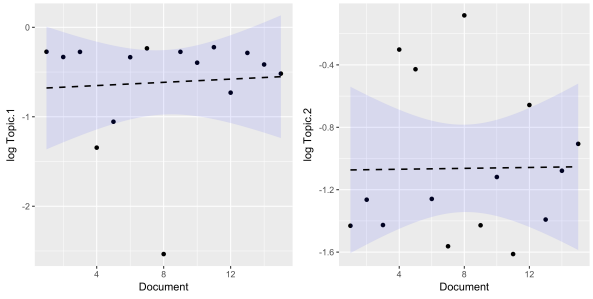

Figure 2: Scatter plot results (log transformation applied to topics) for 2015 document to topic probabilities

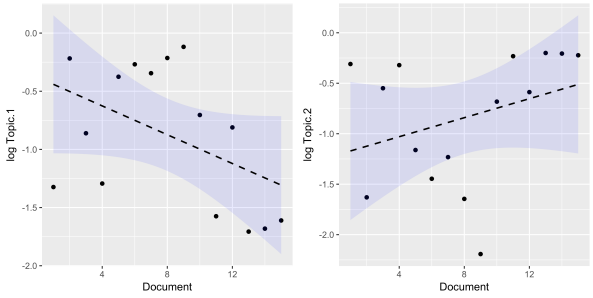

Figure 3: Scatter plot results (log transformation applied to topics) for 2016 document to topic probabilities

The results of the illustration included in Figure 3 appear to have a significant positive or negative linear relationship between the two topics generated by the 2016 data and their associated population of documents. However, the results of the illustration included in Figure 2 do not appear to demonstrate a significant relationship between the generated topic and the population of documents collected in the 2015 data. There is no significant positive or negative linear slope seen between either of the two document to topic probabilities. This result indicates that the two topics generated (See Figure 5) are an invalid representation of the population of documents included in the corpus. When we queried the transcript data in the relational database, it was revealed that one of the documents included in the 2015 sample had 3551 seconds of recorded transcript time from a video titled, “Emerging Technology: The Future of Space”. If we exclude this

document as an outlier, the average seconds of recorded transcript in the 2015 sample was 319.

To confirm the influence of the outlier document included in the 2015 sample, we plotted the density of the data recorded in seconds. This visualization would confirm that that we would need to adjust the value of *K* (number of topics) in order to increase the probability that the LDA algorithm would generate a set of valid topics using this research method.

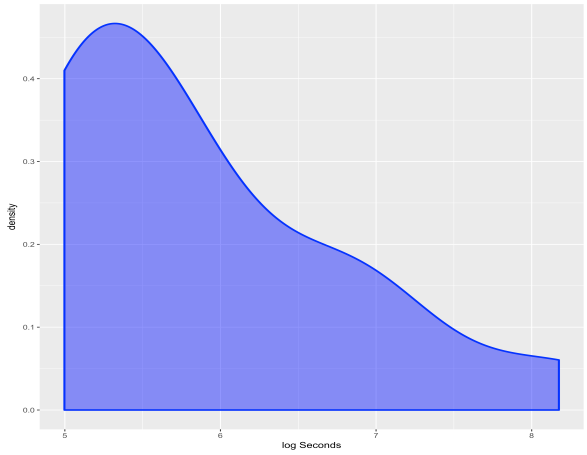

Figure 4: Density plot of 2015 data in seconds

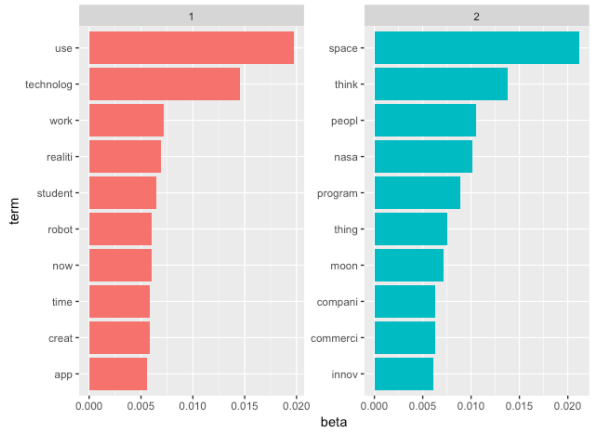

Figure 5: Distribution of terms included in 2015 topic model results, beta = probability distribution

Figure 4 illustrates this heavy skew of transcribed seconds and it is evident that the LDA algorithm was heavily influenced to generate a topic based off of a higher volume of words generated from a specific document or set of documents. Additionally, looking at topic 2, in Figure 5, reveals that there are terms that support the title and content of the outlier document such as "space", "nasa" and "moon". Therefore, we can conclude that the two topics generated by the LDA algorithm are not representative of the entire population of documents collected in the 2015 sample. To remedy this discovery, we would need to increase *K*, or the total number of topics that the LDA algorithm should generate based on the population of documents provided in the sample.

After increasing the number of *K* (topics) in the LDA algorithm to 5, the results told a completely different story for the 2015 sample of documents. Looking at the scatter plot results in Figure 6, we revealed that there was at least one significant topic (topic 1) generated by the algorithm. Topic 1 shows a significant positive linear slope indicating a positive relationship between the population of documents collected in the sample and the generated topic. Therefore, the results of this analysis have indicated that using this research method can assist a researcher in identifying the validity of a topic generated by the LDA algorithm on a corpus of YouTube transcript documents.

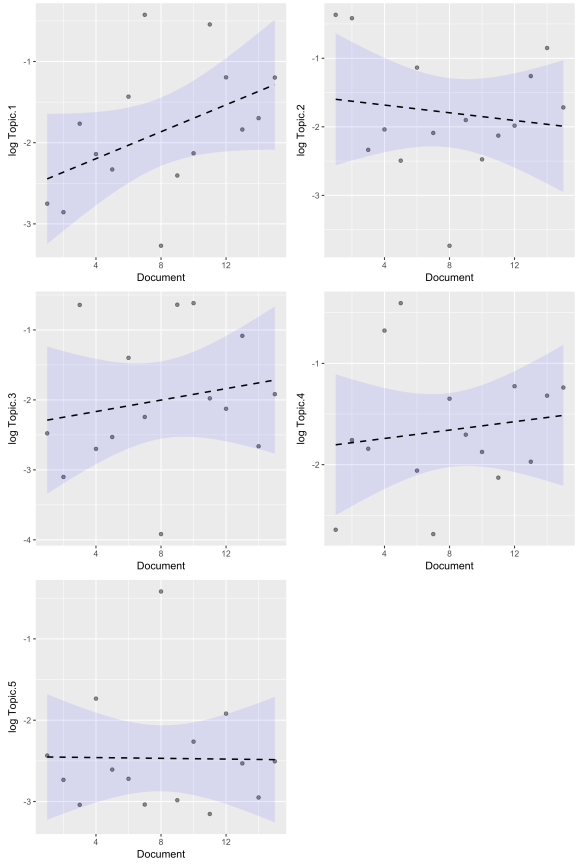

Figure 6: Updated scatter plot results for 2015 document to topic probabilities with 5 topics

Another significant finding discovered from studying the validity of this research method on the YouTube data included our results seen when comparing the output of topics generated by the LDA algorithm with a set of terms defined by an industry hype cycle report. This comparison resulted in a novel

approach to using hype cycles for topic analysis. Specifically, we collected 31 terms that are identified as emerging technologies from the 2015 [11] and 2016 [3] Gartner Hype Cycle for Emerging Technologies reports. These reports cite industry research on the most significant technologies of a given year to provide insight into where an emerging technology fits within the Gartner Hype Cycle over the span of 10 years. The Gartner Hype Cycle defines, through industry research, where a specific technology (in this case – emerging) fits on the cycle by illustrating the expectations of a given technology over time. This progressive cycle includes the following phases: "innovation triggers", "peak of inflated expectations", "trough of disillusionment", "slope of enlightenment", and "plateau of productivity" [18]. Once a technology is placed in the final phase of "plateau of productivity", it is said to be adopted in the mainstream.

We applied a simple approach to comparing the results of the topic models generated using this research method by first creating a corpus of documents that include content from the defined terms found within the Gartner reports. Each emerging technology term found within the Gartner reports include content related to the term name, who provided the analysis of the term, the position and adoption speed justification, user advice, business impact, benefit rating, market penetration, maturity, sample vendors, and recommended reading. All of this content was collectively included in separate documents for a total of 31 documents. The corpus of 31 documents was then imported into the relational database so that an analysis could be performed to compare the Gartner Hype Cycle terms and their supporting metadata with the terms generated by the complete collection of 30 transcript documents from YouTube. Additionally, the published YouTube timestamp from the transcripts collected would fall within the timeframe of the Gartner industry reports thus ensuring that the terms are sensitive to the relative time period. Although the original YouTube transcript documents were collected using search terms such as "emerging technologies", this analysis would further strengthen the validity of the topics generated by the LDA algorithm using this research method and identify them as relevant to industry research in emerging technologies.

Table 5 lists all of the technologies included in the 2015 Gartner Hype Cycle for Emerging Technologies that were individually extracted and transformed to create the corpus of 31 separate documents.

Table 5: Gartner Hype Cycle for Emerging Technologies Report – List of Documents

Technology
4D Printing
802_11ax
Affective Computing
Augmented Reality
Autonomous Vehicles
Blockchain
Brain-Computer Interface
Cognitive Expert Advisors
Commercial UAVs Drones
Connected Home
Context Brokering
Data Broker PaaS_dbrPaaS
Enterprise Taxonomy and Ontology Management
General-Purpose Machine Intelligence
Gesture Control Devices
Human Augmentation
IoT Platform
Machine Learning
Micro Data Centers
Nanotube Electronics
Natural-Language Question Answering
Neuromorphic Hardware
Personal Analytics
Quantum Computing
Smart Data Discovery
Smart Dust
Smart Robots
Software-Defined Anything SDx
Software-Defined Security
Virtual Personal Assistants
Volumetric Displays

We started this comparison by first applying Term Frequency Inverse Document Frequency (TF-IDF) to the corpus of documents containing the Gartner terms. TF-IDF was used on these documents to determine which words included in each document were more favorable than others to use in a SQL query within the relational database [13]. The TF-IDF results were then added to a table within the relational database and compared to the topic modeling results of the LDA algorithm. The topic modeling results from the LDA algorithm included parameters where K (number of topics) = 10 and w (number of words for each topic) = 100. The relationship between the terms generated by the topic model and the Gartner report is considered implicit because each source does not have an explicit relationship with the other. This analysis included results where a specific topic generated by LDA had 15 terms with an implicit relationship with 19 different Gartner terms. Furthermore, this specific topic had a strong relationship, illustrated in Figure 6, with two specific Gartner terms to include "IoT Platform" at 13% of the overall implicit relationship and "Personal Analytics" at 18% respectively. The implicit relationship revealed by this analysis further enforces the value of discovering emerging technology topics

within YouTube that can be acquired using this research method.

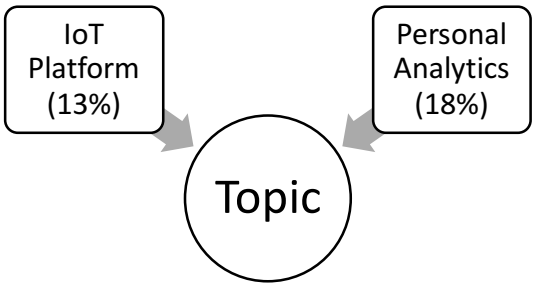

Figure 6: Topic to Gartner Emerging Technology Term Relationship Example

5. Conclusion

Using YouTube transcript data to discover latent topics as a distribution of terms about emerging technologies over slices of time has proven to be a valuable source of information. The discovery of topics in this data requires a unique automated research method that allows for the comparison and analysis of the topics generated by the LDA algorithm. The comparison of topics can be performed with a combination of relational database SQL queries and R analytical packages. Careful consideration must be taken when attempting to interpret the results of the topic analysis.

A set of 30 documents were collected from YouTube video transcripts that include information about "emerging technologies" between 2015 and 2016. Both sets of documents were analyzed separately for comparison and then together in a single corpus. A variety of different visualizations were used to assist with an interpretation of the topic modeling results generated by the LDA algorithm. Results of the analysis demonstrated the need to observe a series of visualizations in a sequence which supports an interpretation that accurately tells a story. The most significant finding of the results was the strategy that can be used to choose which specific topic from the results of the LDA output is valid and should be selected for further interpretation. Specifically, a scatter plot diagram that visualizes the linear relationship between the population of YouTube transcript documents collected and their assigned topic was proven to be a strong indicator for deciding on the value of *K* as input for the LDA algorithm. There is a potential for the results of the topic models generated by the LDA algorithm to have a skew based on the length of seconds transcribed in the YouTube videos. If the length of transcription in the document is too dominant when compared with the other documents in the sample, the LDA algorithm is likely to generate a topic that is heavily skewed toward the dominant document. The value of *K* can then be adjusted by increasing the number of topics generated by the LDA algorithm. This may result in generating topics that have a significant linear distribution of transcript documents across a given topic. We may direct future work towards automating the process for adjusting the value of *K* when executing the LDA algorithm to generate topics.

Finally, the research method used in this study demonstrated significant results for identifying valid latent topics on emerging technologies found within the YouTube data. A sample corpus containing 30 YouTube transcript documents was used to generate 10 topics by the LDA algorithm. These topics were then compared against the emerging technology terms cited in the 2015 and 2016 Gartner "Hype Cycle for Emerging Technologies" reports. The results of this comparison revealed that a topic could be identified as having an implicit relationship with a Gartner emerging technology term. The implicit relationship identified between these two data sources underscores the value of using this research method to discover latent topics on emerging technologies within YouTube.

6. References

[1] C. Cozzens, S. Gatchair, J. Kang, K. Kim, H. Lee, G. Ordonez, and A. Porter, "Emerging technologies: quantitative identification and measurement", Technology Analysis & Strategic Management, Taylor & Francis Group, London, 2010, 22:3, pp. 361-376.

[2] A. Breitzman and P. Thomas, "The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems", Research Policy, Elsevier, Haddonfield, NJ, 2015, 44, pp. 195-205.

[3] M. J. Walker, B. Burton, and M. Cantara, "Hype Cycle for Emerging Technologies, 2016", Gartner, Stamford, CT, 19 July 2016, ID: G00299893, pp. 1-69.

[4] H. van Lente, C. Spitters, and A. Peine, "Comparing technological hype cycles: Towards a theory", Technological Forecasting & Social Change, Innovation Studies, Copernicus Institute of Sustainable Development, Urecht, The Netherlands, 2013, pp. 1615-1628.

[5] YouTube Official Blog: One billion captioned videos, https://youtube.googleblog.com/2017/02/one-billion-captioned-videos.html , accessed on 05-16-2017, published 02-16-2017.

[6] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, 2003, 993-1022.

[7] Selenium with Python Documentation website, http://selenium-python.readthedocs.io/, accessed on 06-08-2017.

[8] A. Chaney and D. Blei, "Visualizing Topic Models", Retrieved from Computer Science at Columbia University: http://www.cs.columbia.edu/~blei/papers/ChaneyBlei2012.pdf, accessed on 05-16-2017, published 06-08-2012.

[9] S. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modeling for qualitative studies", Journal of Information Science, 2017, Vol. 43(1), pp. 88-102.

[10] S. Basu, Y. Yu, V. Singh, and R. Zimmerman, "Videopedia: Lecture Video Recommendation for Educational Blogs Using Topic Modeling", International Conference on Multimedia Modeling, 2016, Springer International Publishing, Miami, Florida, pp. 238-250.

[11] B. Burton and M. Walker, "Hype Cycle for Emerging Technologies, 2015", 07-27-2015, Gartner.

[12] L. AlSumait, D. Barbara, J. Gentle, and C. Domeniconi, "Topic Significance Ranking of LDA Generative Models", Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009, pp. 67-82.

[13] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ.

[14] V. Eckert, C. Curran, and S. Bhardwaj, "Tech breakthroughs megatrend: how to prepare for its impact", PricewaterhouseCoopers, LLP, www.pwc.com/techmegatrend, 2016, accessed on 05-16-2017.

[15] J. Luftman, B. Derksen, R. Dwivedi, M. Santana, H. Zadeh, and E. Rigoni, "Influential IT management trends: an international study", Journal of Information Technology, 2015, pp. 293-305.

[16] T. Douglas, E. Eidam, R. McCauley, B. Miller, T. Harbert, H. Kerrigan, and D. Raths, "Kicking the Tires on Emerging Technology", Public CIO, Technology Leadership in the Public Sector, Winter 2016, pp. 6-10.

[17] F. Alkemade and R. Suurs, "Patterns of expectations for emerging sustainable technologies", Technological Forecasting & Social Change, 2012, Vol. 79, pp. 448-456.

[18] D. O'Leary, "Gartner's hype cycle and information system research issues", International Journal of Accounting Information Systems , 2008, Vol. 9, pp. 240-252.